DOCUMENT RESUME

ED 431 021 TM 029 854

AUTHOR	Swisher, Judith D.; Green, Samuel B.; Tollefson, Nona
TITLE	Using Curriculum-Embedded Assessment for Making Educational Decisions: An Empirical Study with Implications for Including Students with Disabilities in Accountability.
PUB DATE	1999-04-00
NOTE	20p.; Paper presented at the Annual Meeting of the American Educational Research Association (Montreal, Quebec, Canada, April 19-23, 1999).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Accountability; *Decision Making; *Disabilities; Secondary Education; Special Education; *Special Education Teachers; Student Evaluation; *Test Use; Vocational Education
IDENTIFIERS	*Curriculum Embedded Tests; Practical Assessment Exploration System

ABSTRACT

The focus of this research was to examine the perceived effectiveness of the Practical Assessment Exploration System (PAES) (J. Swisher, 1987) for making educational decisions at the instructional program level for students with disabilities across groups of teachers who have various levels of familiarity with the PAES. The PAES is a functional vocational skills curriculum with an embedded assessment of vocational potential. Tasks included in the curriculum are representative of business, home economics, and industrial education, and they provide students with opportunities to complete projects that are simulations of entry-level job tasks. Educators familiar with the PAES completed a survey about its use. Of the responses, 104 were included in the sample, with only 102 included in some analyses. Respondents generally perceived the PAES as more useful for making decisions related to transition planning; aptitude/achievement tests were considered more useful for making decisions associated with general education placements. Responses did not indicate significant differences between the two approaches for making statements about the present level of performance on the student's individualized education plan. Ultimately, the results suggest that students with disabilities are not served well by tests developed for their nondisabled peers. (Contains 7 tables and 33 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 431 021

SCOPE OF INTEREST NOTICE
The ERIC Facility has assigned this document for processing to: TM
In our judgment, this document is also of interest to the Clearinghouses noted to the right. Indexing should reflect their special points of view. EC

Using Curriculum-Embedded Assessment for Making Educational Decisions: An Empirical Study with Implications for Including Students With Disabilities in Accountability

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Judith Swisher
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Judith D. Swisher, Samuel B. Green, Nona Tollefson

University of Kansas

Paper Presentation:

1999 American Educational Research Association Conference - Montreal, Canada

TM029854

BEST COPY AVAILABLE

Using Curriculum-Embedded Assessment for Making Educational Decisions:
An Empirical Study with Implications for Including Students With Disabilities in Accountability

Educational decisions are the heart of accountability issues in standards-based reform. The basic structure that characterizes standards-based reform is grounded in three fundamental decisions: (a) What should students know and be able to do? (b) What do we have to do to get there? and (c) How will we identify the degree to which students have attained the desired knowledge and skills? (Weckstein, 1999). In other words, "We are more likely to end up where we want to go if we clearly identify the destination, focus our efforts on getting there, and check in regularly to make sure we are staying on course. Such a common-sense notion is now supported as well by research, showing that students, including disadvantaged students, can perform at high levels when their education is organized around that framework" (p. 5).

The purpose of this paper is to discuss how the three fundamental decisions influence the education and assessment of students with disabilities and to examine how a curriculum-embedded assessment instrument provides information that assists educators in focusing their efforts and making regular checks to stay on course. More specifically, we investigate the perceived effectiveness of a curriculum-embedded assessment in comparison with traditional assessment measures for evaluating student progress towards curriculum outcomes, documenting the effectiveness of instructional strategies used to help students attain outcomes, and identifying educational and employment related placement options. The particular curriculum-embedded assessment instrument examined in this study has been specifically designed to provide information about how effectively students with disabilities are attaining the knowledge and skills needed to make successful transitions from school to adult life.

A Clearly Defined Destination: What should students know and be able to do?

The decision that students should be prepared to make successful transitions from school to adult life is clearly stated within the 1997 Amendments to the Individuals with Disabilities Education Act (IDEA) (P.L. 105-17). The law requires that students' individual education programs (IEPs) must be designed to prepare students for successful adult outcomes. The law further requires that all students with disabilities: (a) should have the opportunity to participate in the same curriculum that is offered to their non-disabled peers, (b) should participate in the general education curriculum to the maximum extent appropriate for each individual student, and (c) must be included in state and district-wide assessments with the intent of holding schools publicly accountable for their education. Ultimately, each state and local school district must account for the progress of these students towards meeting the same standards set for all students. The theoretical rationale that underscores the inclusion of students with disabilities in the general education curriculum and in state and district assessments is based on the assertion that students who meet state educational standards will be better prepared for successful and productive engagement in the expectations of adult life (Ysseldyke, 1994).

Currently, state standards define what students should know and be able to do within the school setting. Although some states, Kentucky for example, monitor whether students with disabilities make successful transitions from school to adult life, most state standards focus on traditional academic subject areas – reading, writing, math social studies, and science. DeStefano (1993) contended that content standards of this kind do not address the broad-based educational

needs of students with disabilities. "To be well prepared for life after school, some students with disabilities require specific instruction in such areas as general workplace readiness, vocational skills, and independent living skills" (McDonnell, McLaughlin, and Morison, 1997, p.4). Concerns expressed by these and other experts in the field bring post-school transition outcomes center stage for students with disabilities in standards-based reform and accountability (Ginsbert & Berry, 1990; Kortering & Elrod, 1991; McLaughlin, Henderson, & Rhim, 1998).

Focused Efforts For Getting There: What determines effective instructional delivery?

Effective instructional delivery for students with disabilities at the classroom level depends on maintaining an on-going system of assessment that reflects students' educational and transition needs. To the extent that assessment is an on-going and accurate description of achievement, teachers are able to adjust instruction to address students' needs. Much of the assessment information used to develop students' IEPs originates from assessment that is embedded within the curriculum. Moreover, when curriculum-embedded assessment incorporates "authentic" tasks that engage students in the learning/assessment process, students are more likely to be able to demonstrate what they know and can do and "teachers will be able to use the resulting rich information about student learning and performance to shape their teaching in ways that can prove more effective for individual students" (Darling-Hammond, 1994, p. 6).

Assessing the effectiveness of instructional programs for students with disabilities at the district and state level depends on whether assessment results provide accurate descriptions of what students know and are able to do and whether the information is sufficiently used to identify program improvement needs (Weckstein; 1999). According to Weskstein (1999) large-scale assessment reform calls for measures that hold schools accountable and, at the same time, inform and enhance the instructional process. Thurlow, Elliott, Ysseldyke, and Erikson (1996) argued that assessment and instruction should be viewed as inextricably linked and that the results of district and state assessments should be presented in a manner that provides useful information to those who need it for instructional purposes.

Many large-scale assessments, however, have placed teachers in a no-win situation with pressures to boost achievement scores on tests that have been "designed specifically to fulfill an accountability function rather than an instructional function" (Popham, 1998, p.4). Popham's argument is clearly aimed at the inadequacy of using standardized tests alone for assessing the quality of education and identifying instructional program needs. Thurlow and Ysseldyke (1993) cautioned that large-scale assessments should not be limited to a single assessment format and that assessment developers should explore ways of obtaining comparable measures from alternative forms of assessment. Moreover, Olsen and Ysseldyke (1999) recommended a number of more subjective classroom-based measures as alternate assessment options for students with disabilities who cannot participate in the regular state assessments even with accommodations. Currently, the Kentucky alternate portfolio used to assess students with disabilities who do not participate in the regular state assessment includes teacher's instructional data as one measure of student performance (Kearns, Kleinert, & Kennedy, 1999). Consequently, there is a need to examine the viability of including curriculum-embedded assessments in conjunction with more traditional measures for accountability.

Regular Checks To Stay On Course: Using curriculum-embedded measures for accountability

While curriculum-embedded assessments are increasingly recognized as valuable sources of information for on-going classroom assessment and internal supports for school-based inquiry (Darling-Hammond, Ancess, & Falk, 1995; Roeber, 1996; Swisher & Green, 1998; Wolf & Baron, 1996), questions remain about the utility of classroom-based assessment for large-scale assessment. The use of curriculum-embedded assessment formats in addition to traditional measures for large-scale assessment is at least implied by Title I of the Elementary and Secondary Education Act. Currently, states that receive Title I funds are required to include "multiple measures" in their state assessments. Gribbons, Sheinker, Carlson, and Winter (1998) contend that while "there is no statutory definition, 'multiple measures' can be thought of as falling along a continuum that ranges from multiple item or task types on a single assessment instrument through multiple instruments incorporating a variety of formats" (p.1). While various statistical procedures have been recommended for combining results from different assessment formats (Ryan, Martios, Winter, & Gribbons, 1998), if curriculum-embedded scores are to be combined with other assessment formats for accountability decisions, it is critical that technically sound evaluation processes and scoring rubrics are used to score curriculum-embedded assessments.

According to Kopriva (1998), the assessment of students who learn, process, and respond differently requires the standardization of constructs and processes as opposed to the standardization of specific responses in order to obtain accurate and comparable assessment results for all students. Kopriva further contended that the use of scoring rubrics in large-scale testing has demonstrated the viability of standardizing constructs and processes to achieve comparable results. Similarly, Popham has contended that, "Rubrics used to assess skills in large-scale assessments should not be task-specific, that is, designed to score responses to only a specific task. Rather, instructionally useful rubrics must be skill-focused, that is, designed to evaluate responses to any task representing the skill" (Popham, 1998, p. 8). If curriculum-embedded assessments are based on appropriately operationalized processes and scoring rubrics, they have the potential for providing assessment information that holds schools accountable and, at the same time, informs and enhances the instructional process.

A Larger Question: Feasibility Issues and the Accountability / Instruction Balance

At the policy level, feasibility issues such as higher costs and time constraints present additional factors that affect the accountability / instruction balance. While educators and researchers tend to agree that large-scale assessment should hold schools accountable, provide information that improves instruction, and include multiple assessment formats so that a more accurate description of student achievement will emerge (Darling-Hammond, 1994; Roeber, 1996; Weckstein, 1999), decision-makers at the policy level are faced with the dilemma of available resources and the amount of time required to conduct alternative assessments for state and local accountability.

Cost and time constraints are important feasibility concerns for large-scale accountability (Roeber, 1996). A growing number of researchers and educators, however, tend to view feasibility from a more theoretical perspective resulting in questions that address the underlying purposes and the conceptual framework of accountability. Educators and researchers including Fredricksen (1984), Darling-Hammond (1994), Kearns, Kleinert, and Kennedy (1999), and

Popham (1997) argue that the accountability / instructional trade-off is not a necessary consequence of large-scale assessment. While Gardner and Hatch (1989) argued that "huge amounts of money have been invested in standard psychometric instruments whose limitations have become increasingly evident" (p. 109), Fredricksen (1984) argued that the higher costs of using alternative forms of assessment could be justified if one of the primary purposes of assessment was to improve instruction. The questions now asked are: Is it "feasible" to continue to use traditional tests as the primary measure for accountability when these tests do not adequately inform and improve teaching and learning? Is it productive to continue to separate the accountability process from classroom instruction and curriculum implementation? Is it truly cost effective for large-scale accountability systems to continue to measure student progress at the expense of guiding and improving instructional opportunities?

According to studies conducted by Boyer (1983), Goodlad (1984), and Sizer (1985), the use of standardized tests during the initial accountability efforts of the 1970s, had negative effects on teaching and learning in high schools. Similarly, Darling-Hammond (1994) argued against school reform strategies that use assessment as a lever for external control of schools suggesting that the effects of improper use and application of basic skills tests have been "most unfortunate for the students they were most intended to help [disadvantaged students and those with disabilities]. . .Thus, the quality of education made available to many students has been undermined by the nature of the testing programs used to monitor and shape their learning" (p.12). She further argued that accountability efforts that rely on external control of schools are "unlikely to be successful and the assessments are unlikely to be equitable because they stem from a distrust of teachers and fail to involve teachers in the reform process . . . Teachers understandings of students' strengths, needs, and approaches to learning are not well supported by external testing programs that send secret, secured tests into the school and whisk them out again for machine scoring that produces numerical quotients many months later" (p. 5-6).

The potential lack of relevance of traditional tests to future employability raises yet another feasibility issue associated with the over reliance on traditional tests for accountability, especially for students with disabilities. Studies conducted by Eckland (1980), Gordon & Sum (1988), and Jaeger (1991), as cited by Darling-Hammond (1994), revealed that student scores on basic skills tests are not related to employability or job-related earnings. In response to the need for relevant measures of employment capability, a study conducted by Swisher and Green (1998) compared a curriculum-embedded assessment - the Practical Assessment Exploration System (PAES) - and two traditional aptitude tests for predicting job-related outcomes obtained three to five years later for students with disabilities. The two traditional tests were the Career Ability Placement Survey (CAPS) (Knapp, Knapp, & Knapp-Lee, 1981/1992) and the Differential Aptitude Test (DAT) (Bennett, Seashore, & Alexander, 1973/1992). The curriculum-embedded measure, the PAES, was most strongly related to the job-related outcome that measured level of support required on a job, but also tended to be related to the other two criteria, salary and hours worked. The CAPS and the DAT were almost uniformly very weakly related to the job-related outcomes.

Objectives of the Study

Given the predictive capability of a curriculum-embedded assessment – the PAES – (Swisher & Green, 1998), the need for all students with disabilities to be included in state and

district assessments, and the need for state assessments to include multiple measures, we were interested in examining the usefulness of curriculum-embedded assessment, specifically the PAES, for making educational decisions for students with disabilities. The objectives of the present study were (a) to compare teachers' perceptions of the usefulness of the PAES, traditional aptitude/achievement tests, and interest/employability inventories for making educational decisions typically made for students with disabilities (e.g., present level of performance statements for the IEP, goals and objectives for the IEP, functional skill needs, job placements, vocational class placements, level of support needed on a job, and support needed for a vocational class) and (b) to examine differences in the perceived usefulness of the PAES across groups of teachers with different levels of familiarity with the PAES. Before presenting the methodology and results of our study, we will provide a brief overview of the PAES.

The Practical Assessment Exploration System

The Practical Assessment Exploration System (PAES) (Swisher, 1987) is a functional vocational skills curriculum with an embedded assessment of vocational potential. The conceptual framework of the PAES, as illustrated in Figure 1, is based on features of various types of alternative assessments including: performance-based, authentic, dynamic, and curriculum-embedded assessment where assessment tasks and exploration tasks are the same.

Tasks included in the curriculum are representative of three contexts – business, home economics and industrial education – and provide students with opportunities to solve problems and complete projects that are simulations of actual tasks performed on entry-level jobs. The categories for each context are presented in Table 1. Each category includes six tasks which are designed to increase in level of difficulty from one task to the next. For example, the first of the six tasks associated with alphabetizing involves placing 26 cards, one for each letter of the alphabet, in alphabetical order. The sixth alphabetizing task involves placing 117 cards in alphabetical order. The sequential nature of the tasks allows students to apply what they learn from one task to another. In this way the PAES tasks are useful learning tools as well as measures of specific skills (Swisher & Green, 1998).

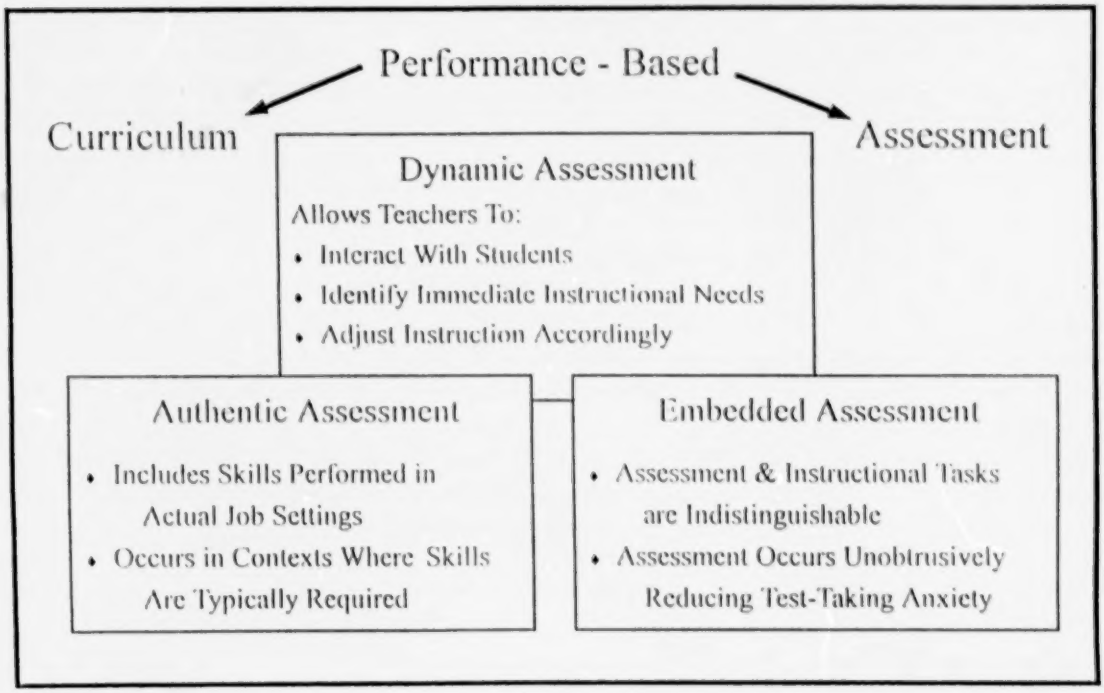
Table 1

Type of Tasks for the Three PAES Contexts

Business	Home economics	Industrial education
Alphabetical filing	Liquid and dry measurement	Linear measurement tools
Filing title, author, and date	Food preparation by recipes	Wrenches and bolts
Filing by numerical sequence	Basic food service tasks	Hammers and screwdrivers
Collating papers	Food scale	Hand saws
Making change	Measuring cloth	Electrical wiring projects
Operating a cash register	Sewing by hand	Sheet metal projects
Operating a ten-key calculator	Using a sewing machine	Wood projects
Creating a data base	Cloth construction	
Word processing		

Figure 1.

The PAES - Conceptual Framework


Performance - Based
Curriculum
Assessment
Dynamic Assessment
Allows Teachers To:
• Interact With Students
• Identify Immediate Instructional Needs
• Adjust Instruction Accordingly
Authentic Assessment
• Includes Skills Performed in Actual Job Settings
• Occurs in Contexts Where Skills Are Typically Required
Embedded Assessment
• Assessment & Instructional Tasks are Indistinguishable
• Assessment Occurs Unobtrusively Reducing Test-Taking Anxiety

Authentic assessment principles are integrated throughout the PAES implementation process. As suggested by Wiggins (1993) and Gardner (1993) tasks represent "messy" real-world contexts and offer "authentic-simulations" of typical job situations. In addition, assessment procedures allow students to have access to resources and accommodations, such as: charts, written procedures, diagrams, and other reference materials that would normally be made available on a job. As students are actively engaged in a variety of authentic tasks that capitalize on student interest, there is reason to expect that the tasks serve to stimulate a positive work ethic: perseverance, self-motivation, high standards, and self-confidence (Darling-Hammond et al., 1995). Finally, the authentic nature of the PAES class allows teachers the opportunity to identify behaviors that would potentially interfere with successful performance in vocational classes or entry-level jobs (Swisher and Clark, 1991).

The dynamic and embedded-assessment properties of the PAES allow for on-going interactions between the student and the teacher where the teacher provides only the amount of assistance that is necessary for students to complete the tasks. Students are provided with verbal, gesture, and actual demonstration assistance in graduated increments so that they receive assistance only when absolutely necessary. As students grow accustomed to a pattern of leading questions, hints, and prompts that offer strategies for answering their own questions, they tend to make a greater effort to think for themselves, thus requiring less assistance from the teacher. This scaffolded process of assistance provided by the teacher identifies the type and amount of support needed by the student to complete each task and serves to provide a clear description of the student's instructional needs (Swisher & Green, 1998). Students typically work in the PAES class one or two hours a day for as many as eighteen weeks or more. This experience provides on-going feedback on student performance using multiple measures taken over time. According to Herman, Aschbacher and Winters (1992) and Gipps (1994) repeated assessment, over time and across a range of contexts allows the teacher to build a more comprehensive understanding of a student achievement.

The assessment process is operationalized through a series of steps that students and teachers follow for each task. The PAES results describe student potential by evaluating skills associated with work independence, accuracy, and speed. Scoring rubrics are "skill-focused" as opposed to "skill-specific" (Popham, 1998). For example, separate rubrics are used to rate students on five criteria: (a) amount of assistance required to complete each task, (b) quality of performance on the first trial, (c) work rate, (d) the number of trials it takes to complete a task correctly from the beginning to the end, and (e) level of interest for each task The teacher rates the student on each criteria for each task. The scores are then collapsed across tasks for each criteria and ultimately aggregated to produce overall scores.

Method

The present study used survey methodology to ask educators familiar with the PAES, achievement/ aptitude tests, and employability/interest inventories to rate the usefulness of the three types of assessments for making educational decisions for students with disabilities. The questionnaire also asked respondents a series of questions that provided information for grouping individuals according to their level of familiarity with the PAES and the other two measures. Initial contact letters were sent to identify schools that would be willing to participate in the study. Questionnaires were sent and follow-up contacts were made at three weeks and five weeks.

Research Participants

Of the group of 150 schools that had the PAES at the time of the study, 17 schools had only recently obtained the PAES and 10 schools were no longer using the system. Of the 123 eligible schools, 77% or 90 schools responded to the initial contact letter that asked whether they would be willing to participate in the study and to distribute questionnaires to four individuals in the school. Of the 360 questionnaires that were sent to the 90 schools, 44% or 160 questionnaires were returned representing 55 schools. Of these 160 educators, 104 had sufficient data to be included in the sample. In some analyses the sample size was reduced to 102 due to the requirement that all data be present to conduct the repeated-measures analyses. The majority of respondents were females, who have graduate level education, special education teacher certification, high school experience primarily with students who have learning disabilities, mild to moderate retardation, and/or behavior disorders. Approximately one half of the respondents were from metropolitan areas. Over one-third of the respondents were from rural areas.

Familiarity Groups

Respondents were grouped according to their level of familiarity with the PAES. Three groups were formed - high, moderate, and low familiarity. Respondents in the low familiarity group (N = 45) rated themselves as either unfamiliar or moderately familiar with the PAES. The low familiarity group also had to meet at least one of the following criteria: (a) had not recorded PAES assessment data, (b) were unfamiliar with the PAES Summary Report, (c) knew that the PAES Summary Report had never been used in their district, (d) did not know if the report had ever been used, or (e) never had training from a PAES representative on how to administer the PAES. In addition to these qualifiers, respondents in the low familiarity group either did not spent any time during their work day taking students through the PAES or had not been involved with the PAES for more than a year.

In order to be included in the high familiarity group (N = 25), respondents had to rate themselves as very familiar with the PAES. They also had to meet at least one of the following criteria: (a) had recorded the PAES assessment data, (b) were familiar with the PAES Summary Report, (c) knew that the PAES Summary Report had been used in their district, or (d) had received training from a PAES representative to administer the PAES. In addition to these qualifiers, respondents in the high familiarity group had to have spent more than 25% of their day taking students through the PAES or had to have been involved with the PAES more than one year.

Respondents who did not meet the criteria for the low familiarity group or the high familiarity group were included in the moderately familiar group (N = 32). It is possible for persons in the moderately familiar group to have rated themselves as unfamiliar, moderately familiar, or very familiar with the PAES. Persons in the moderately familiar group were excluded from the low familiarity group if (a) they either had some knowledge of the PAES assessment procedures and reports or (b) they had spent a reasonable amount of time being involved with the PAES. In contrast, persons in the moderately familiarity group were excluded from the high familiarity group if (a) they had no knowledge of the PAES assessment procedures and reports or (b) they had not spent a sufficient amount of time involved with the PAES.

Survey

The questionnaire asked respondents to rate the three measures for their usefulness in making educational decisions related to transition planning and development of the student's IEP. Respondents were also asked to indicate the extent to which they had: (a) used the PAES in their school district, (b) been trained in administering the PAES and interpreting the PAES results, (c) been involved in making various decisions associated with the IEP and transition planning, (d) were familiar with three types of assessments (e.g., the PAES, achievement and aptitude tests, interest and employability skills inventories), and (e) found information from the three types of assessments useful for making various decisions associated with IEP development and transition.

Analyses

Only scores for respondents who considered themselves to be at least moderately familiar with aptitude/achievement tests and interest/employability inventories were included in the analyses. Differences in the perceived usefulness of the three measures were examined by conducting a three-way ANOVA with two within-subjects factors (type of test and type of decision) and one between-subjects factor (level of familiarity with the PAES). Level of familiarity with the PAES had three levels: (a) low familiarity, (b) moderate familiarity, and (c) high familiarity. Type of test had three levels: (a) the PAES, (b) achievement/aptitude tests, and (c) interest/ employability inventories. Type of decision had nine levels associated with planning, placement, and support. The nine levels were: (a) entry-level job placements, (b) vocational class/training placements, (c) general education class placements, (d) present level of performance statements for the IEP, (e) goals and objectives for the IEP, (f) goals and objectives for transition plans (g) functional skills that need to be developed, (h) type and amount of support needed for an entry-level job, (i) type and amount of support needed for a vocational class. The dependent variables were Likert scale scores that indicate the extent of usefulness of each type of test for making the nine decisions.

A general linear model analysis was also conducted to investigate whether responses across the nine types of decisions vary as a function of level of familiarity. The general linear model analyses included not only the factors in the previous ANOVA but also familiarity as a quantitative predictor. Of particular interest was the interaction of familiarity and the other factors.

Results

Differences in the Perceived Usefulness of the Three Measures

A three-way ANOVA was conducted with type of test and type of decision as two within-subjects factors and level of familiarity with the PAES as the between-subjects factor. Two of the three main effects were significant: type of tests, Wilk's $\Lambda = .53$, $F(2,98) = 44.29$, $p < .001$, multivariate $\eta^2 = .48$; and type of decisions, Wilk's $\Lambda = .47$, $F(8,92) = 2.75$, $p < .001$, multivariate $\eta^2 = .53$. Two of the two-way interactions were significant: type of test by level of familiarity with the PAES, Wilk's $\Lambda = .85$, $F(4,196) = 4.10$, $p = .003$, multivariate $\eta^2 = .08$; and type of test by type of decision effect, Wilk's $\Lambda = .32$, $F(16,84) = 11.01$, $p < .001$,

multivariate η^2 = .68. All other sources were non-significant. Because the two main effects for type of tests and type of decisions were factors that were involved in the two significant two-way interactions, these main effects will not be interpreted.

Follow-up Analyses for Type of Test by Level of Familiarity with the PAES

Follow-up analyses for the interaction between type of test and level of familiarity with the PAES were conducted to examine differences in the perceived usefulness of the three types of tests –the PAES, aptitude/achievement tests, and interest/ employability inventories – for each level of familiarity – high familiarity, moderate familiarity, and low familiarity. Three one-way ANOVAs, one for each level of familiarity, were conducted as follow-up analyses to examine the simple main effects of the three types of tests for each level of familiarity averaging across the nine decisions. Testing at a .05 alpha level, analyses yielded significant results for each familiarity group: low familiarity, Wilk's Λ = .62, F (2,24) = 7.47, p = .003, multivariate η^2 = .38; moderate familiarity, Wilk's Λ = .58, F (2,31) = 11.17, $p < .001$, multivariate η^2 = .42; and high familiarity, Wilk's Λ = .37, F (2,43) = 37.03, $p < .001$, multivariate η^2 = .63. These results indicate significant differences in the perceived usefulness of the three types of tests for each familiarity group. In all cases, the PAES had the highest means. See Table 2 for means and standard deviations.

Follow-up Paired Sample T-test Comparisons Among Types of Tests for Each Familiarity Group

Nine paired-sample t-tests were conducted as follow-up tests to the simple main effect one-way ANOVAs to examine mean differences among the three types of tests within each level of familiarity. Using Bonferroni to control for Type I error, mean differences for all comparisons involving the PAES with the other two types of tests were significant at .0167 for each familiarity group. Comparisons for aptitude/achievement tests and interest/employability inventories were non-significant for all levels of familiarity. Table 3 presents results for all significant comparisons.

An inspection of Table 2 should allow us to understand why we obtained a familiarity by type of test interaction. The difference between the mean for the PAES and the means for the other two types of tests is a greater for the high familiarity group than for the low and moderate familiarity groups. These results indicate that persons who are more familiar with the PAES tend to perceive the PAES as generally more useful for making decisions than the other two types of tests.

Follow-up Analyses for Type of Test for Each Type of Decision

Follow-up analyses for the type of test by type of decision interaction were conducted to examine differences in the perceived usefulness of the three types of tests – the PAES, aptitude/achievement tests, and interest/employability inventories – for making each of the nine decisions averaging across familiarity groups. Nine follow-up analyses, one for each of the nine decisions, were conducted to examine the simple main effects of the three types of tests for each of the nine decisions across level of familiarity. Using Bonferonni testing at a .005 alpha level, analyses yielded significant results for all nine decisions. These results indicated significant differences in the perceived usefulness of the three types of tests for each of the nine decisions.

Table 4 reports the results for these analyses. See Table 5 for means and standard deviations for each type of test for each of the nine decisions. The PAES had the highest means in comparison with the other two types of tests on eight of the nine decisions. Aptitude/achievement tests had the highest mean for general education placement decisions. It is the differential pattern of means for general education placements that produced the test by familiarity interaction.

Follow-up Paired Sample T-tests Among Type of Tests for Each Decision

Follow-up analyses to the nine significant one-way ANOVAs were conducted to examine differences among the three types of tests for each of the nine decisions averaging across familiarity groups. Twenty-seven paired sample t-tests were conducted. Controlling for Type I error across the twenty-seven tests using Bonferroni, $p < .0018$, seven of the nine comparisons involving the PAES and aptitude/achievement were significant. In all seven comparisons the PAES was considered to be more useful in making decisions than aptitude/achievement, with two exceptions. As expected achievement/aptitude tests were considered as more useful than the PAES for making general education placement decisions. The second exception was in developing IEP level of performance statements where there was no significant difference between the PAES and achievement/aptitude tests. Seven of the nine comparisons involving the PAES and interest/employability inventories were significant. In all seven comparisons the PAES was considered to be more useful in making decisions than interest/employability inventories with two exceptions. As expected both tests were considered equally useful for making job placement decisions. The second exception was in making general education placement decisions where both tests were also considered to be equally useful. The results of the paired-sample t-tests for comparisons involving the PAES and the other two tests are presented in Table 6. Table 7 presents the means and standard deviations for each level of familiarity for each decision.

Discussion

The challenges of including all students with disabilities in state and district assessments have been addressed in this paper within a framework of contemporary arguments that explore and support the following issues: (1) Students with disabilities will be better prepared for the expectations of adult life if state standards include transition related outcomes, (2) At the classroom level students are more likely to be able to demonstrate what they know and can do when curriculum-embedded assessments incorporate "authentic" tasks that engage students in the learning/assessment process, (3) The use of traditional tests alone for accountability does not adequately assess the quality of education, identify program needs, or serve to enhance teaching and learning, (4) Feasibility issues beyond cost and time should be considered when assessments prove to be inadequate measures of student achievement or when they fail to improve instruction, and (5) The potential of combining curriculum-embedded assessment results with other assessment formats for accountability decisions is a viable option only if technically sound evaluation processes and scoring rubrics are developed and used so that teachers can make judgements that are useful across dimensions of learning and consistent across schools.

In this study, we were interested in examining the perceived usefulness of a curriculum-embedded assessment – the PAES – for making educational decisions at the instructional program level for students with disabilities across groups of teachers who have various levels of familiarity

with the PAES. With only a few exceptions results suggest that teachers, regardless of their familiarity with the PAES, prefer using the PAES to aptitude/achievement tests and interest/employability inventories for making decisions associated with IEP development, transition planning, employment and vocational training. Moreover, the perceived usefulness of the PAES tended to increase with level of familiarity with the PAES.

It was not surprising that the PAES was perceived as more useful for making decisions related to transition planning and that aptitude/achievement tests were considered as more useful than the PAES for making decisions associated with general education placements. The PAES was specifically designed to assess transition needs rather than specific academic skills. It was somewhat surprising that results did not indicate significant differences between the PAES and aptitude/achievement tests for making present level of performance statements on the IEP considering the predominately academic nature of many such statements. One possible explanation is that 87% of the individuals who responded to the survey indicated that special education teachers in their schools systematically receive the PAES assessment results. Since, in most cases, these teachers are responsible for writing present level of performance statements for the IEP, it is possible that they have found the PAES assessment results to be an equally useful indicator of overall strengths and needs.

Ultimately, these results suggest that students with disabilities are not served well by tests developed for their non-disabled peers. Results also suggest a need to explore the potential of using assessments embedded within the curriculum for these students to assess dimensions of learning that are more academically focused. More specifically, it is important to explore how this could be accomplished within large-scale assessment across core academic domains so that a more accurate description of student achievement will emerge.

There are many unanswered questions about how curriculum-embedded assessment could be put in practice in academic areas so that teachers endorse the process and gain sufficient knowledge to make judgements that are consistent across schools. While Calfee and Hiebert (1988) have argued that teachers do assess students, collect data, and make decisions that influence educational programs, they also contend that there is a need to enhance the technical quality of the process. Furthermore, in order to maintain high standards with less standardization, teachers will require staff development that will enable them to evaluate and eliminate sources of unfair bias in the scoring of instructionally embedded assessments, balance subjectivity and objectivity, use their subjective knowledge of students appropriately in selecting tasks and assessment options while adhering to common, collective standards of evaluation (Darling-Hammond, 1994). The consequence of sufficient staff development is that "students will learn more as a result of assessment, rather than being more precisely classified, and schools will be able to inquire into and improve their practices more intelligently, rather than being more rigidly ranked" (p. 18). Finally, the use of curriculum-embedded assessment in large-scale accountability as one of "multiple" assessment formats, potentially "cases teachers in the role of problem framers and problem solvers who use their classroom and school experiences to build an empirical knowledge base to inform their practice and strengthen their effectiveness" (p.26).

Table 2

Means and Standard Deviations for Type of Test for each Familiarity Group

	High familiarity			Moderate familiarity			Low familiarity		
Test	M	SD	n	M	SD	n	M	SD	n
PAES	5.09	.50	45	4.80	.71	33	4.51	.79	26
Apt / Achiev tests	3.48	1.19	45	4.11	1.04	33	3.91	.85	26
Int / Employ inventories	3.37	1.41	45	3.93	1.16	33	3.78	.99	26

Note. The sample sizes for the high, moderate, and low familiarity groups were 25, 32, and 45, respectively.

Table 3

Paired Sample T-Test Comparisons among Types of Tests for Each Familiarity Group

	Paired differences					
	The PAES vs apt/ach tests			The PAES vs int/emp inventories		
Familiarity group	Mean diff	SD diff	t-value	Mean diff	SD diff	t-value
High familiarity	1.61	1.34	8.10***	1.72	1.42	8.16***
Moderate familiarity	.69	1.11	3.57**	.87	1.04	4.80***
Low familiarity	.60	.88	3.45**	.74	1.00	3.76**

*p < .05 **p < .01 ***p< .001

Table 4

Follow-up Analyses for Type of Test for Each Type of Decision

Decision	Wilk's Λ	F-value	p-value	Eta^2
Job placement	.53	44.27	.000	.48
Vocational class	.59	33.52	.000	.41
General education placement	.72	19.39	.000	.28
IEP present level of performance	.62	30.96	.000	.38
IEP goals and objectives	.67	24.70	.000	.33
Transition planning	.63	29.09	.000	.37
Functional skill needs	.39	78.80	.000	.61
Support needs on a job	.46	59.39	.000	.54
Support needs in a vocational class	.43	66.82	.000	.57

Note. Degrees of freedom for all analyses were 2 and 100, respectively.

Table 5

Means and Standard Deviations for Type of Test for each Type of Decision

		The PAES		Aptitude/ achievement tests		Interest/ employability inventories	
Decision	n	M	SD	M	SD	M	SD
Job placements	102	4.93	1.18	3.47	1.40	4.67	1.30
Vocational class placements	102	5.09	1.18	3.70	1.42	4.48	1.30
General education placements	102	3.01	1.89	4.25	1.59	3.03	1.66
IEP present level of performance	104	4.68	1.32	4.29	1.50	3.10	1.88
IEP goals and objectives	104	4.97	1.11	4.31	1.38	3.74	1.75
Transition plans	104	5.26	.94	3.98	1.50	4.36	1.68
Functional skill needs	104	5.50	.78	3.61	1.71	3.13	1.89
Support needs on a job	104	5.16	1.05	3.18	1.72	3.24	2.00
Support needs in a vocational class	104	5.13	1.00	3.26	1.60	3.13	1.94

Table 6

Paired-sample T-tests among Type of Tests for Each Decision

	Paired differences					
	The PAES vs apt/ach tests			The PAES vs int/emp inventories		
Decision	Mean diff	SD diff	t-value	Mean diff	SD diff	t-value
Job placements	1.46	1.63	9.07 ***	.26	1.65	1.62
Vocational class placements	1.38	1.57	8.87 ***	.61	1.41	4.36 ***
General education placements	-1.24	2.57	-4.85 ***	.02	1.89	.10
IEP present level of performance	.38	1.87	.039	1.59	2.06	7.85 ***
IEP goals and objectives	.66	1.55	4.37 ***	1.23	1.68	7.47 ***
Transition plans	1.28	1.60	8.13 ***	.90	1.71	5.37 ***
Functional skill needs	1.89	1.86	10.39 ***	2.38	1.93	12.54 ***
Support needs for a job	1.98	1.92	10.52 ***	1.92	2.04	9.63 ***
Support needs in a vocational class	1.88	1.72	11.14 ***	2.01	1.91	10.71 ***

*p < .05 **p < .01 ***p< .001 .

Table 7

Means and Standard Deviations for Each Level of Familiarity with the PAES and Each Decision

	High familiarity		Moderate familiarity		Low familiarity	
	Mean	SD	Mean	SD	Mean	SD
Job placements	5.13	.99	5.00	1.32	4.32	1.38
Vocational class Placements	5.40	.89	5.09	1.12	4.52	4.50
IEP present level of performance	5.02	1.03	4.58	1.50	4.23	1.42
Transition plans	5.49	.73	5.33	.92	4.77	1.11
Functional skill Needs	5.69	.56	5.58	.56	5.08	1.13
Support needs for a job	5.44	.89	4.88	1.14	5.04	1.11
Support needs in a vocational class	5.33	.95	5.09	.91	4.85	1.12

Note. The sample sizes for the high, moderate, and low familiarity groups were 25, 32, and 45, respectively.

References

Bennett, G. K., Seashore, H. G., & Alexander, G. W. (1992). Differential aptitude tests: Technical manual (5th ed.). San Antonio, TX: Harcourt Brace, The Psychological Corporation.

Boyer, E. L. (1983). High school. New York: Harper & Row.

Calfee, R., & Hiebert, E. (1988). The teacher's role in using assessment to improve learning. In E. Freeman (Ed.). The Forty-Eighth Educational Testing Service Invitational Conference, (pp. 45-61). Princeton, NJ: Educational Testing Service.

Darling-Hammond, L. (1994). Performance-based assessment and educational equity. Harvard Educational Review, 64(1), 5-29.

Darling-Hammond, L, Ancess, J., & Falk, B. (1995). Authentic assessment in action. Studies of schools and students at work. New York, NY: Teachers College Press.

DeStefano, L., (1993). The effects of standards and assessment on students in special education (Synthesis Report 10). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Eckland, B. K. (1980). Sociodermographic implications of minimum competency testing. In R. M. Jaeger & C. K. Tittle (Eds.) Minimum competency achievement testing: Motives, models, measures, and consequences (pp. 124-135). Berkeley, CA: McCutchan.

Frederiksen, N. (1984). The real test bias. influences of testing on teaching and learning. American Psychologist, 39(3), 193-202.

Gardner, H., & Hatch, T. (1989). Multiple intelligences go to school. Educational Researcher, 18(8), 104-110.

Ginsbert, R. & Berry, B. (1990). Experiencing school reform: The view from South Carolina. Phi Delta Kappan, 71, 549-552.

Goodlad, J. I. (1984). A place called school: Prospects for the future. New York: McGraw-Hill.

Gordon, B., & Sum, A. (1988). Toward a more perfect union: Basic skills, poor families, and our economic future. New York: Ford Foundation.

Gribbons, B., Sheinker, A., Carlson, D., & Winter, P. (1998). Towards a research agenda for multiple measures: Joint CCSSO/CRESST research plan (based on a proposal written for CCSSO's State Collaborative on Assessment and Student Standards). Council of Chief State School Officers.

Jaeger, R. M. (1991). Legislative perspectives on statewide testing: Goals, hopes, and desires. Paper presented at the American Educational Research Association Forum, Washington, DC.

Kearns, J. F., Kleinert, H. L., & Kennedy, S. (1999). We need not exclude anyone. Educational Leadership, 56(3), 33-38.

Knapp, L., Knapp, R. R., & Knapp-Lee, L. (1992). Technical manual: CAPS career ability placement questionnaire. San Diego, CA: EDITS.

Kopriva, R. (1998). LEP, special education, and Title I large-scale assessment - can we find a one-size-fits all approach that works? Paper presented at the National Conference on Large Scale Assessment, Colorado Springs, CO.

Kortering, L. J., & Elrod, G. F. (1991). Programs for adolescents with mild handicaps: Evaluating where we are and contemplating change. Career Development for Exceptional Individuals, 14(2), 145-157.

McDonnell, L. M., McLaughlin, M. J., & Morison, P. (1997). Educating one & all: Students with disabilities and standards-based reform. Washington, DC: National Academy Press.

McLaughlin, Henderson, & Rhim, (1998). Snapshots of reform: How five school districts are interpreting standards based reform for students with disabilities. Alexandria, VA: Center for Policy Research on the Impact of General and Special Education Reform.

Popham, W. J. (1997). What's wrong - and what's right - with rubrics. Educational Leadership, 54(10), 72-75.

Popham, W. J. (1998). Where large scale educational assessment is heading and why it shouldn't (Opening session address 28th Annual National Conference on Large Scale Assessment). Colorado Springs, CO: Council of Chief State School Officers.

Roeber, E. D., (1996). Designing coordinated assessment systems for Title I of the improving America's schools act of 1994 (Fourth in a series of papers on the promises and challenges of IASA Title I). Washington, DC: Council of Chief State School Officers.

Ryan, J., Martios, J., Winter, P., & Gribbons, B. (1998). Towards a research agenda for multiple measures. Paper presented at the National Conference on Large Scale Assessment, Colorado Springs, CO.

Sizer. T. (1985). Horace's compromise. Boston: Houghton Mifflin.

Swisher, J. D. (1987). The Practical Assessment Exploration System. Jacksonville, FL: Talent Assessment, Inc.

Swisher, J. D., & Green, S. B. (1998). An empirical study comparing curriculum-embedded assessment and traditional aptitude measures for predicting job-related outcomes for students with disabilities. Educational Assessment, 5(1), 57-70.

Thurlow, M. L., Elliott, J., & Ysseldyke, J. E., Erikson, R. (1996). Questions and answers: Tough questions about accountability systems and students with disabilities. (Synthesis Report 14). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thurlow, M. L., & Ysseldyke, J. E. (1993). Can "all" ever really mean "all" in defining and assessing student outcomes? (Synthesis Report 5). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Weckstein, P. (1999). School reform and enforceable rights to quality education. In J. Heubert (Ed.), Law and school reform: Six strategies for providing educational equity: New Haven, CT: Yale University Press.

Wolf, & Baron, (1996). Performance-based student assessment: Challenges and possiblilities. 95th Yearbook of the National Society for the Study of Education, Part I. Chicago, IL: University of Chicago Press.

Ysseldyke, J. (1994). Students with disabilities & educational standards: Recommendations for policy & practice. NCEO Policy Directions No. 2. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Ysseldyke, J., & Olsen, K. (1999). Putting alternate assessments into practice: What to measure and possible sources of data. Exceptional Children, 65(2), 175-185.